# Social Role Discovery in Human Events

Vignesh Ramanathan* Bangpeng Yao† Li Fei-Fei†
*Department of Electrical Engineering, Stanford University
†Computer Science Department, Stanford University
{vigneshr, bangpeng.yao, feifeili}@cs.stanford.edu

## Abstract

*We deal with the problem of recognizing social roles played by people in an event. Social roles are governed by human interactions, and form a fundamental component of human event description. We focus on a weakly supervised setting, where we are provided different videos belonging to an event class, without training role labels. Since social roles are described by the interaction between people in an event, we propose a Conditional Random Field to model the inter-role interactions, along with person specific social descriptors. We develop tractable variational inference to simultaneously infer model weights, as well as role assignment to all people in the videos. We also present a novel YouTube social roles dataset with ground truth role annotations, and introduce annotations on a subset of videos from the TRECVID-MED11 [1] event kits for evaluation purposes. The performance of the model is compared against different baseline methods on these datasets.*

## 1. Introduction

Humans are social animals. Our ability to comprehend human relations stands fundamental to our survival, development and social life. We understand such relationships in terms of social roles assumed by people, and tend to describe events using these roles. For instance, we would describe the birthday video in Fig. 1 as "Parents helping the birthday boy cut a cake", rather than "Two people helping another person cut a cake". Typically, social roles answer semantic queries like, "Who is doing what in an event?". While the tasks of identifying the action and detecting the person are widely studied in computer vision, the problem of role assignment is relatively new and equally interesting.

Social role discovery derives motivation from the field of "Role Theory" [2] in sociology, which observes that people behave in predictable ways based on their social roles. This shows that knowing the role of a person can help determine his/her interactions with the environment and vice-versa. In computer vision, [13] leveraged the same intuition to build a human activity recognition model. Also, the knowledge of social roles can help determine the interesting segments of social event footages [7] and sports videos.

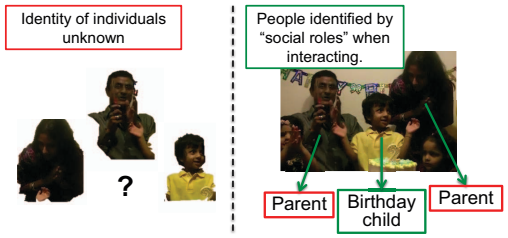



Figure 1. When people interact in an event, they assume event specific social roles. Social roles act as identities for the individuals and can help us describe the event in terms of these roles. Role recognition is fundamental in understanding a human event.

The definition of social roles is event specific, and can sometimes be abstract such as, people "helping", "visiting" or "residing" in a nursing home [13], making role identification a difficult human task. Ideally, we would like to automatically discover such interaction-based role assignments in any event. Also, annotating roles is time consuming and needs knowledge of the event. Recognizing these difficulties, we formulate the problem of social role discovery in a weakly supervised framework. Given a set of videos belonging to a social event without training labels for the people in the videos, we group them into different social roles. The event label acts as the weak annotation in our setting, restricting the discovered roles to be event specific.

The problem is amply challenging due to the wide variation in appearance, scale, location and scene context of a role across different videos as seen in Fig. 2. As illustrated in Fig. 1, it is difficult to determine roles by observing people individually. Rather, social role discovery is an attempt to identify people based on their interactions in an event. Modeling such interactions in the absence of role labels during training acts as an additional challenge.

In order to solve this problem of weakly supervised role assignment, we propose a Conditional Random Field (CRF) to capture inter-role interaction cues, and develop


Figure 2. Sample frames from different events in the YouTube Social Roles dataset are shown with ground truth role annotations used for evaluation. The different roles in each event are marked by the colors noted in the last column. The huge variation in appearance, location, scale and scene context for a role across different videos can be seen.

a tractable variational inference procedure to jointly learn role labels as well as model weights. Further, to evaluate the model performance, we introduce a novel YouTube social roles dataset in Sec. 5.1, accompanied by event specific ground truth role annotations for the people in the videos. It is to be noted that the role labels are only used for model evaluation and not for the training. We also provide role annotations for a subset of videos from two events of the TRECVID MED-11 [1] event kits, and test our model performance on these videos. Experiments on these datasets show that our method achieves encouraging performance in weakly supervised social role assignment.

## 2. Related Work

**Socially aware video and image analysis** Recent works on social network construction and interaction understanding is relevant to our work on social role recognition. [25] associates people in a video using face recognition and track matching. [4, 5] clusters people in a movie into adversarial groups. [5] uses scene context and visual concept attributes to build social relation network. [23] also builds a social role network based on their co-occurrence of movie characters in different scenes. These works do not group people across different videos, but consider people within one movie. [22] uses appearance features to predict the relationship between people by training on images with weak relationship labels, while [19] performs occupation classification based on clothing and context in human images. [20] studied the problem of face recognition in social context.

**Social Interaction in Action Recognition** Another related line of work has been the use of social interaction to aid group action recognition [14, 3, 6]. [14] explicitly models human interaction, while [3] uses features of people in spatio-temporal vicinity to detect group activities and jointly track multiple people. [18] also uses social grouping to help multi target tracking. [10] uses social context in group photos to make better prediction of human attributes and scene semantics. [9] recognizes group social activities through attribute learning. [17] develops interaction features based on facial orientation to recognize activities like hand-shaking. Similarly, [16] also models facial attention. Although the above works capture social interactions in some form, they do not explicitly identify the roles assumed by people during a social event.

**Role Recognition** Recently, [7, 13] used social roles to predict group activities. [7] found face attention patterns in first person videos to detect interaction activities like monologue, discussion and dialogue. They clustered faces in training videos based on attention patterns, and represented frame sequences by histogram of cluster occurrences. [13] predicted role labels like "defender" and "attacker" in sports videos to identify group activities. They used training labels

to learn role assignments based on spatio-temporal interaction between players. However, in our work we are not provided role annotations, and we wish to discover interaction-based roles automatically by studying different instances of an event. We also use richer interaction features.

## 3. Our Approach

We define social role discovery as a weakly supervised problem, where the training role labels for the people in the videos are not available. We are only provided the event label for each video, and the number of roles to be discovered in an event. We assume that every video is pre-processed to obtain individual human tracks similar to [6, 13].

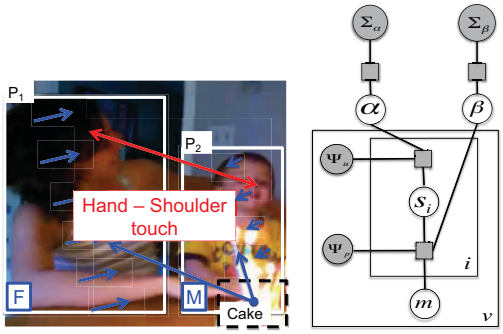
Social roles are not only decided by person specific descriptors, but also by the interaction between people. Hence, any model used to discover social roles should be capable of incorporating this information. However, interaction in an event is usually restricted to a small set of roles. In our approach, every event has a reference role, and the interaction of any person with this reference role is most significant. To understand this, consider a *birthday*, where the important interactions mostly involve the "birthday person". With this assumption, it is sufficient to model the interaction of any person only with the reference role. This is a realistic simplification, enabling us to perform tractable inference as shown in Sec. 4. One instance of the reference role is assumed to be present in every video belonging to the event class. We refer to the other roles as secondary roles.

### 3.1. Model Formulation

We present a CRF model which accounts for the reference role interaction with other roles in a video. An overview of our approach is shown in Fig. 3, along with the factor graph of our model. As illustrated, to capture person specific social cues, we extract unary features ($\Psi_u$) from each human track, describing spatio-temporal activity, human appearance and human-object interaction. Similarly, to represent interaction based social cues, pairwise features ($\Psi_p$) describing proxemic touch codes, and spatial proximity are extracted. Our CRF model uses these features to perform weakly supervised social role recognition.

Let $\mathbb{P}_v$ be the set of people in a video $v$ and $s_i^v$ be the social role assigned to a person $p_i^v \in \mathbb{P}_v$. We want to assign social roles, and jointly learn model weights by maximizing the log likelihood of the CRF shown in Eq. 1.

$$\underset{s_E,\alpha,\beta}{\operatorname{argmax}} \sum_v \Bigg\{ \sum_{p_i^v} \alpha \cdot \Psi_u(p_i^v, s_i^v) + \sum_{p_j^v \neq p_m^v} \beta \cdot \Psi_p(p_m^v, p_j^v, s_j^v) - Z_v \Bigg\} - \frac{\alpha^T \Sigma_\alpha^{-1} \alpha + \beta^T \Sigma_\beta^{-1} \beta}{2}, \tag{1}$$




(a) (b)

Figure 3. (a) The features extracted by our model are illustrated on a sample birthday video frame. Unary features are represented in blue, while the pairwise features are shown in red. (b) The factor graph of our CRF model is shown. The observed variables are shaded. $m$ is the index of the reference role in the video $v$. The model variables are as defined in Sec. 3.1.

where $m_E$ denotes the reference role in the event $E$, and $p_m^v$ the person holding the reference role in $v$. The model potentials are defined as

$$\begin{aligned} \alpha \cdot \Psi_u(p_i^v, s_i^v) &= \sum_s \alpha_s \cdot \mathbf{1}(s = s_i^v) \Psi_u(p_i^v), \\ \beta \cdot \Psi_p(p_m^v, p_j^v, s_j^v) &= \sum_{s \neq m_E} \beta_s \cdot \mathbf{1}(s = s_j^v) \Psi_p(p_m^v, p_j^v) \end{aligned} \tag{2}$$

In Eq. 1, $s_E$ is the complete social role assignment to all people in the event, and $Z_v$ is the log-partition function for the video $v$. $\Sigma_\alpha$ and $\Sigma_\beta$ are the covariances of the Gaussian priors on $\alpha$ and $\beta$ respectively. Note that the model only considers interaction of different roles with the reference role, in accordance with our assumption, and every video is assumed to contain one person playing this reference role. $\alpha$ and $\beta$ are the unary and pairwise weights to be learnt respectively. A factor graph of the model is shown in Fig. 3

### 3.2. Unary Features

The unary feature $\Psi_u$ captures role specific social cues extracted from human tracks, and their interaction with the event environment. $\Psi_u$ can be expanded into four components as shown below.

**Histogram of Gradient Feature** $\Psi_u^{HoG}$: Bag of densely computed HoG3D [11] words of dimension 1429 along the human track is used as low-level features to capture the individual actions.

**Spatio-Temporal Feature** $\Psi_u^{ST}$: A person's movement in an event is another useful cue regarding his/ her role. For example, the "bride" often walks down the aisle in a church *wedding*. The human motion between two frames is

binned along 8 directions to form a trajectory feature similar to [12]. These features are normalized across different people in a video to partly account for camera motion.

**Object Interaction Feature** $\Psi_u^{OI}$: The interaction of a person with the event environment plays a key role in determining his/ her role. "birthday person" cutting a "cake" and "function host" talking at the "lectern" are representative examples. In the current work, we extract interaction features corresponding to only these two objects in the respective events. [8] is used to obtain specific object detection scores in a video. These scores are spatially pooled similar to [15] in the periphery of the person's bounding box and averaged across multiple frames to form an object interaction feature of dimension 48 for every event object.

**Social Feature** $\Psi_u^{Soc.}$: These features capture two important social aspects of a person, representing gender and clothing. Such cues are important in events like *wedding*. This would also capture the gender bias in certain roles like "brides". We first use [27] to detect faces, and obtain scores [1] for gender classification. The scores are averaged across frames to form the gender feature. The clothing of a person is represented by the RGB color histogram with 32 bins.

### 3.3. Pairwise Interaction Features

Human interaction forms an important basis for social role definitions. For instance, the "parent" in a *birthday* is distinguished from "guests" by their interaction with the "birthday person". Similarly "bride-groom", "instructor-student" interactions separate the respective roles from others. These interactions are recorded by the pairwise feature $\Psi_p$ composed of two components as shown below.

**Proxemic Interaction Feature** $\Psi_p^{Prox.}$: The proxemic interaction of two people provides interesting insights regarding the relation between roles in an event such as the touch-code between a "parent" and the "birthday child". The use of proxemics for describing human-human relations was introduced in [24], where the authors classify proxemics between two people into 6 classes with 20 models. Proxemics are also referred as touch-codes, indicating the way people touch each other. For every pair of humans in a video, we use all 20 models from [24] to find proxemic scores in different frames. The scores are normalized across all human pairs in a given video and split into 16 bins for every model, to form our final proxemic descriptor. The scores are set to a minimum value, if a pair of people are never sufficiently close to each other.

**Spatio-Temporal Interaction Feature** $\Psi_p^{ST}$: The spatial separation of people across time is a simple but powerful measure of human interaction in a video. For instance, the "bride" and "groom" are always near each other in a wedding, while the "groomsmen" are farther away from the "bride". The spatial distance between a pair, normalized by bounding box dimensions at different time instants are used.

[1] We use software from http://cmp.felk.cvut.cz/ fisarond/demo/

## 4. Inference

The difficulty of solving Eq. 1 arises due to the correlation between different social roles and the coupling introduced by $Z_v$. [26] proposed a mean field approximation to solve Conditional Topic Random Fields, with simple chain connected CRFs and CRFs without interaction potentials. Along similar lines, we develop a variational inference method to find an approximate solution for our graphical model. We show that the simplifying assumption of interactions being restricted to the reference role, helps us perform tractable inference as a part of the optimization procedure. We also introduce a variational approximation to the social role probability distribution in a video, with similar dependencies as the original model.

We formulate the variational approximation $q$ of the model distribution as shown in Eq. 3, where $s_v$ denotes the role assignment to all people in the video $v$.

$$q(\alpha, \beta, s_E|\lambda_\alpha, \lambda_\beta, \sigma_\alpha^2, \sigma_\beta^2, \phi, \psi) = \prod_j q(\alpha^j|\lambda_{\alpha^j}, \sigma_{\alpha^j}^2) \prod_k q(\beta^k|\lambda_{\beta^k}, \sigma_{\beta^k}^2) \prod_v q(s^v|\phi^v, \psi^v) \tag{3}$$

The distributions over $\alpha$ and $\beta$ are approximated by univariate normal distribution with means given by $\lambda_\alpha, \lambda_\beta$ and variances $\sigma_\alpha^2, \sigma_\beta^2$. $\phi^v$ is a factor giving the probability of a person being assigned the reference role in the video. $\psi^v$ is a set of $|\mathbb{P}_v|$ factors, where $\psi_{(i)}^v$ is the secondary role probability matrix for other people in the video, when $p_i^v$ is assigned the reference role. $\phi, \psi$ are formally defined in Eq. 4. This variational approximation of the social role probability, retains the dependencies in our original structure. It represents one predominant reference role, with secondary role assignments dependent on this reference role.

$$\begin{aligned} \phi^v(p_i^v) &= p(s_i^v = m_E) \\ \psi_{(i)}^v(p_j^v, s) &= p(s_j^v = s|s_i^v = m_E),\ j \neq i,\ s \neq m_E \end{aligned} \tag{4}$$

Inference is then carried out through coordinate ascent. In each iteration, the updates for $\phi, \psi$ require inference in the CRF model, with the model weights fixed. When the model weights are fixed, our graph reduces to a tree for each individual video, allowing us to perform exact clique-tree inference. The optimization procedure and update equations for $\psi, \phi, \lambda, \sigma^2$ are shown in the supplementary document Sec. A, due to space limitations.

We initialize both $\phi^v$, $\psi_{(i)}^v$ to be uniform for all people in the event. $\lambda_{\alpha_s}$ are initialized to be the maximally separated points in the unary feature space for an event $E$. $\lambda_{\beta_s}$

are similarly initialized from the pairwise interaction feature space. $\sigma^2_{\alpha^j}$ are initialized to $0.01$ or $0.1$ based on the variance of the event unary features. Similarly, $\sigma^2_{\beta^k}$ are initialized to $10$ or $0.1$ for all events.

In every video $v$, the person $p^v_m$ with the highest value of $\phi^v$ is assigned the reference role, forming a reference role cluster. The corresponding variational probability $\psi^v_{(m)}$ is used to assign secondary roles to other people in the video. We enforce a lower and upper bound on the number of people assigned to a secondary role cluster in the event. In practice, the bounds are set to a $10\%$ range of the smallest and largest ground-truth cluster sizes in the event. This acts a lose prior on the number of people in each role cluster. Linear integer programming is used to satisfy these constraints during role assignment, whose details are shown in supplementary document Sec. B due to space limitations.

# 5. Experiment and Results

## 5.1. Datasets

**YouTube Social Roles** Most publicly available video datasets are not suitable for evaluating the social role assignment task, since they do not cover a good range of people donning different roles in specific social events. In an attempt to evaluate our method, we collected a set of YouTube videos under $4$ social events. The details of the dataset are shown in Tab. 1. To facilitate easy evaluation, we annotate every person in our dataset with the relevant social roles. Some videos have stray individuals not annotated with any specific social role and are called as "others". Again it is to be noted that role labels are used only for evaluation.

Within each social event, there is wide variation in event settings as seen from the sample video frames in Fig. 2. *Wedding* and *Birthday* videos were chosen to cover both indoor and outdoor celebrations. *Award ceremony* includes graduation functions, presidential award functions as well as corporate events. Similarly, *physical training* refers to martial arts, aerobics and other forms of fitness classes. This diversity in scenarios, with the same underlying interactions between different roles is an interesting characteristic of the dataset, and makes the task amply challenging.

**TRECVID Social Roles** Among publicly available datasets, the TRECVID-MED11 event kits [1] have two social event classes *birthday* and *wedding*. However, most of the videos in these kits either have very few characters or crowd activities where people cannot be distinguished from each other. Hence, we chose a smaller subset, covering reasonable number of people in different roles. Some videos were cropped to include only the parts showing relevant social events. Details of the dataset are shown in Tab. 2

Since human tracking is not the focus of the current work, we obtain human tracks through the active learning tool from [21]. The dataset along with the human tracks, and role annotations would be made publicly available [2].

| Event Name | Social Roles (No. of people per role) | No. of videos | Avg. duration |
|---|---|---|---|
| Birthday Party | birthday child (40), parents (44), friends (71), guests (28) | 40 | 80.84 sec. |
| Catholic Wedding | bride (40), groom (40), priest (38), grooms men (45), brides maids (43), others (8) | 40 | 88.74 sec. |
| Award Function | presenter (40), receipient (309), host (25), disributor (17), others (13) | 40 | 111.13 sec. |
| Physical Training | instructor (36), students (127) | 36 | 50.49 sec. |

Table 1. Details of the YouTube social roles dataset.

| Event Name | Social Roles (No. of people per role) | No. of videos | Avg. duration |
|---|---|---|---|
| Birthday Party | birthday person (34), parent/spouse (40), friends (59), guests (31) | 34 | 44.65 sec. |
| Catholic Wedding | bride (34), groom (34), priest (29), grooms men (29), brides maids (29) | 34 | 72.00 sec. |

Table 2. Details of the TRECVID social roles dataset.

## 5.2. Role Discovery Results

In our experiments, we evaluate the model by comparing results with human annotated roles in each video. Due to the weakly supervised nature of the problem, we do not have a direct mapping between role clusters and ground-truth role labels. To facilitate easy comparison with different baselines, the role clusters obtained from a method are each mapped to one of the human defined roles, maximizing the total correct role assignments in an event. We present results on the two datasets from Sec. 5.1 and compare our full model against different baselines in Tab. 3, 4. The tables show the total accuracy of role assignment in an event. The baselines used for comparison are explained below.

- prior: Simple baseline. A random person in each video is assigned the reference role, and the true prior of secondary roles is used to assign roles to other people in the video.
- k-means: Simple experiment, where people are clustered using appearance and spatio-temporal features.
- CRF with $\Psi_u$: To judge the importance of interaction features, we use a CRF with only unary features, similar to CTRF in [26]. We use same priors as our model.
- CRF with $\Psi_{up}$: To demonstrate the gain in modeling inter-role interactions, instead of using interactions as context, the mean interaction feature of a person with

[2] https://sites.google.com/site/eevignesh/socialroles

| Method | Birthday | Wedding | Award Function | Physical Training |
|---|---|---|---|---|
| prior | 29.32% | 20.17% | 62.97% | 65.93% |
| k-means cluster | 33.88% | 29.43% | 31.97% | 57.67% |
| CRF with $\Psi_u$ | 38.25% | 39.22% | 69.31% | 76.69% |
| CRF with $\Psi_{up}$ | 41.53% | 38.83% | 77.75% | 77.91% |
| Our model - $\Psi_p^{Prox.}$ | 43.72% | 36.41% | 79.54% | **82.82%** |
| Our model - $\Psi_p^{Spat.}$ | 43.72% | 39.32% | 79.80% | 77.91% |
| Our Full Model | **44.81%** | **42.72%** | **83.12%** | **82.82%** |

Table 3. Total role assignment accuracy for the YouTube dataset. The best performance in each event is marked by bold font.

| Method | Birthday | Wedding |
|---|---|---|
| prior | 28.72% | 21.63% |
| k-means cluster | 29.88% | 34.19% |
| CRF with $\Psi_u$ | 35.98 % | 38.71 % |
| CRF with $\Psi_{up}$ | 42.07% | 41.94 % |
| Our model - $\Psi_p^{Prox.}$ | 41.46% | 41.29% |
| Our model - $\Psi_p^{Spat.}$ | 43.90% | 41.29% |
| Our Full Model | **44.51%** | **43.87%** |

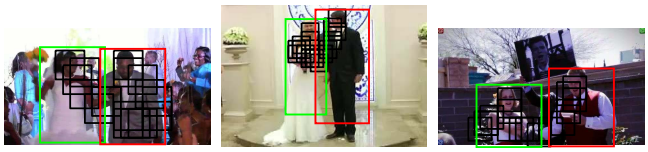
Table 4. Total role assignment accuracy for the TRECVID dataset. The best performance in each event is marked by bold font.

everyone else is concatenated with the unary feature, forming $\Psi_{up}$, used in same CRF as before.

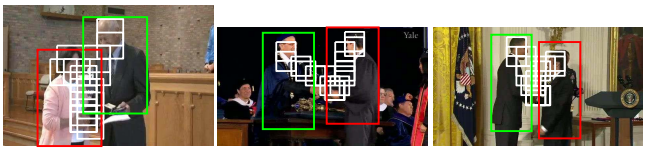
- Our model - $\Psi_p^{Prox.}$: Full model without $\Psi_p^{Prox.}$
- Our model - $\Psi_p^{ST}$: Full model without $\Psi_p^{ST}$

From results in Tab. 3, we notice that a CRF using $\Psi_u$ outperforms naive k-means clustering, justifying the use of this representation with our unary features. Also, the use of interaction as a context feature in $\Psi_{up}$ is seen to do better than the use of only unary features, in most events. This confirms our belief that, human interactions are informative for role recognition. In particular, we observe a considerable increase for the *award function* event, where the interaction between the "recipient" and "presenter" as seen in Fig. 4(b) would help distinguish the "presenter" from other people at the dais. Next, we observe that our full model shows significant improvement over CRF with $\Psi_{up}$. This demonstrates the value in explicitly modeling interaction between role pairs, instead of using interaction as a context feature. For instance, consider a wedding with similar interactions between a "bride-groom" pair, and a "bridesmaid-groomsman" pair. These interactions lead to the same interaction-context feature, for both the "bride" and the "bridesmaid". However, our full model would treat them differently, due to the difference in the other role participating in the interaction, leading to a richer description.

Our full model using the complete pairwise interaction feature $\Psi_p$ performs better than the models only using $\Psi_p^{Prox.}$ or $\Psi_p^{ST}$, showing the gain from use of both the components. It is interesting to note the considerable drop in performance for *ward function* and *wedding* events, in the absence of $\Psi_p^{Prox.}$. We observed that the proxemic models corresponding to specific touch-codes fired consistently across different "bride-groom" and "presenter-recipient" pairs in *wedding* and *award functions* respectively, distinguishing them from other role pairs in the events. We illustrate this in Fig. 4.



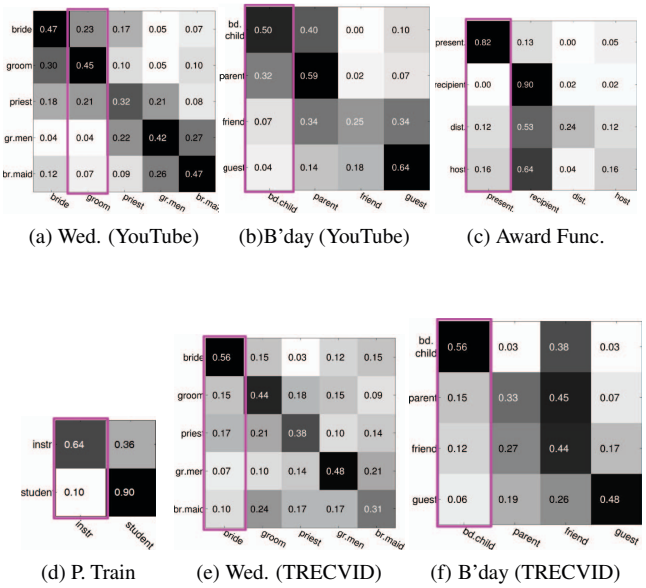
(a) Wedding



(b) Award function

Figure 4. Sample frames from videos are shown, where our full model identified the correct (a) "bride" (green box), "groom"(red box) roles in *wedding* and (b) "presenter" (green box), "recipient" (red box) roles in *award function*. The same Hand-Hand touch code is seen to be detected on different instances of the same role pair. The black and white boxes are the part detections from two different proxemic models for Hand-Hand touch.

To analyze the complete role assignment, we look at the confusion matrices in Fig. 5. The column corresponding to the reference role cluster chosen by our algorithm is highlighted in each matrix. The average purities of the reference role clusters are 0.65 and 0.56, in the YouTube and TRECVID datasets respectively. This demonstrates the ability of our model to isolate the reference role in each video. We observe that the model is able to cluster the roles better in the *wedding* event, as seen in Fig. 5(a), 5(e). This can be accounted to the strong interaction between the "bride" and "groom", separating them from the remaining roles. To study this interaction, we visualize the marginals of the spatial relationship of different roles with the reference role ("groom") cluster in the YouTube *wedding* dataset, in Fig. 6. The marginals capture the expected interaction, as explained in the figure. The confusion of "distributor" with the "recipient" in Fig. 5(c), can be explained by the similar patterns of interaction between the "recipient" receiving the award from the "presenter", and the "distributor" handing out the award to the "presenter". "friends" are difficult to distinguish from "guests" in the TRECVID *birthday* dataset, where we observed both roles to exhibit low interaction with the reference role.




(a) Wed. (YouTube) (b)B'day (YouTube) (c) Award Func.

(d) P. Train (e) Wed. (TRECVID) (f) B'day (TRECVID)

Figure 5. Confusion matrices for different events are shown for the YouTube and TRECVID Social Roles dataset. The column corresponding to the reference role cluster chosen by our model is highlighted in each event. (This figure is best viewed in color)

Sample results from our full model are shown in Fig. 7 along with typical failure instances. Most failure cases involved less interaction among people, as seen in the last column of *birthday*, *wedding* and *physical training*.

In order to evaluate the latent reference role assignment in our model, we compare performances with a control setting which randomly chooses the reference role in each video. The average accuracy of role assignment over all events is seen to drop by $4.82\%$ for the YouTube social roles dataset with this choice of reference role, justifying the need to model it as a latent variable. In particular, we observe a large drop of $6.80\%$ for the *wedding* event, which has more role classes than the other events leading to increased randomness in the choice of reference role in each video.

## 6. Conclusion

We proposed to recognize social roles from human event videos in a weakly supervised setting, and designed a CRF to model the inter-role interactions along with person specific unary features. This weak supervision enables our method to automatically understand the relations between people, and discover the different roles associated with an event. It further reduces the human effort involved in observing long video footages to annotate the roles. We showed considerable performance improvement over different baseline models. As a next step, our approach can be extended to perform simultaneous event classification along with role discovery. It is also noted that our method is not robust to noisy and fragmented reference role tracking, due to the inherent assumption of one reference role per video. In the future, we wish to account for such noisy tracking.

## Acknowledgements

We thank A. Alahi, J. Krause and K. Tang for helpful comments. This research is partially supported by the DARPA-Mind's Eye grant, and the IARPA-Aladdin grant.

## References

[1] Trecvid multimedia event detection track. 1, 2, 5

[2] B. J. Biddle. Recent development in role theory. *Annual Review of Sociology*, 12:67–92, 1986. 1

[3] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In *ECCV*, 2012. 2

[4] L. Ding and A. Yilmaz. Learning relations among movie characters: A social network perspective. In *ECCV*, 2010. 2

[5] L. Ding and A. Yilmaz. Inferring social relations from visual concepts. In *ICCV*, 2011. 2

[6] C. Direkolu and N. OConnor. Team activity recognition in sports. In *ECCV*. 2012. 2, 3

[7] A. Fathi, J. K. Hoggins, and J. M. Rehg. Social interactions: A first person perspective. In *CVPR*, 2012. 1, 2

[8] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. on PAMI*, (9), 2010. 4

[9] Y. Fu, T. Hospedales, T. Xiang, and S. Gong. Attribute learning for understanding unstructured social activity. In *ECCV*, 2012. 2

[10] A. C. Gallagher and T. Chen. Understanding images of groups of people. In *CVPR*, 2009. 2

[11] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008. 3

[12] A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*. 2011. 4

[13] T. Lan, L. Sigal, and G. Mori. Social roles in hierarchical models for human activity recognition. In *CVPR*, 2012. 1, 2, 3

[14] T. Lan, Y. Wang, W. Yang, S. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. *IEEE Trans. on PAMI*, 34(8):1549–1562, 2012. 2

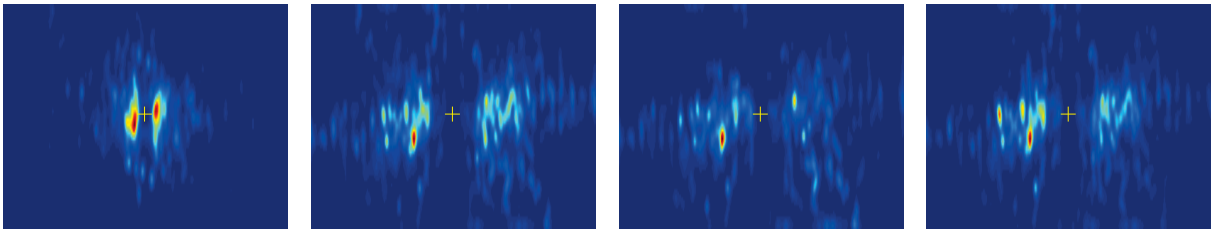
[15] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010. 4

[16] M. Marin-Jimenez, A. Zisserman, and V. Ferrari. "heres looking at you, kid"-detecting people looking at each other in videos. In *BMVC*, 2011. 2

[17] A. P. Perez, M. Marszalek, A. Zisserman, and I. Reid. High five: Recognising human interactions in tv shows. In *BMVC*, 2010. 2

[18] Z. Qin and C. R. Shelton. Improving multi-target tracking via social grouping. In *CVPR*, 2012. 2

[19] Z. Song, M. Wang, X. Hua, and S. Yan. Predicting occupation via human clothing and contexts. In *ICCV*, 2011. 2



(a) Bride (b) Priest (c) Bridesmaids (d) Groomsmen

Figure 6. Marginal of the position of a role relative to the reference ("groom"), estimated by our model is shown for YouTube *wedding* videos. The spatial positions in the image co-ordinates are normalized by width of the reference role bounding box, and binned. The groom's position is marked by a cross-hair. The "bride" is mostly close to the "groom". "groomsmen" and "bridesmaids" are distributed around the groom as expected. The uncertainty in recognizing the "priest" is reflected by a scattered distribution.

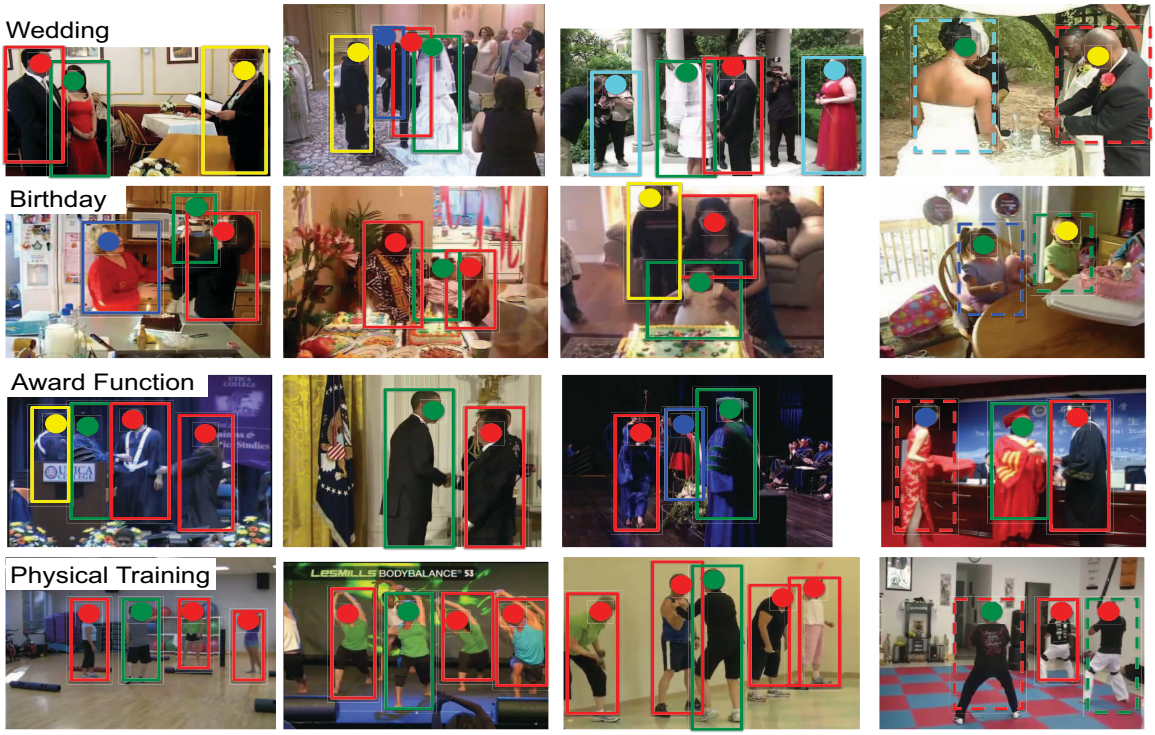



Figure 7. Sample results from the YouTube social roles dataset is shown, where each row corresponds to an event. Boxes with solid lines indicate correct role assignments from our full model, while dashed lines represent faulty assignments. Different roles are indicated by the same color code as in Fig. 2. The ground truth role of a person is indicated by the color of the dot on the person. Last column shows typical failure cases for each event.


[20] Z. Stone, T. Zickler, and T. Darrell. Toward large-scale face recognition using social network context. In *Proc. of the IEEE*, 2010. 2

[21] C. Vondrick and D. Ramanan. Video annotation and tracking with active learning. In *NIPS*, 2011. 5

[22] G. Wang, A. Gallagher, J. Luo, and D. Forsyth. Seeing people in social context: recognizing people and social relationships. In *ECCV*, 2010. 2

[23] C.-Y. Weng, W.-T. Chu, and J.-L. Wu. Rolenet: Movie analysis from the perspective of social networks. *IEEE Trans. on Multimedia*, (2):256–271, 2009. 2

[24] Y. Yang, S. Baker, A. Kannan, and D. Ramanan. Recognizing proxemics in personal photos. In *CVPR*, 2012. 4

[25] T. Yu, S.-N. Lim, K. Patwardhan, and N. Krahnstoever. Monitoring, recognizing and discovering social networks. In *CVPR*, 2009. 2

[26] J. Zhu and E. P. Xing. Conditional topic random fields. In *ICML*, 2010. 4, 5

[27] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012. 4